Codierung von Vereins- und Verbändenamen mit INTEXT

Klein, Harald

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Klein, H. (1996). Codierung von Vereins- und Verbändenamen mit INTEXT. *Historical Social Research*, *21*(3), 146-153. https://doi.org/10.12759/hsr.21.1996.3.146-153

Historical Social Research, Vol. 21 — 1996 — No. 3, 146-153

METHODS: SOFTWARE, REVIEWS, NOTICES

Codierung von Vereins- und Verbändenamen mit INTEXT

*Harald Klein**

Abstract: This article is meant to demonstrate how text data are converted into numeric data. In a research project concerning parlamentarian leading groups it was necessary to produce new data files from already existing files. There were data files, e.g. one of the deputies of the Weimar Reichstag and another that contained informations about the deputies of the German Bundestag 1995, which though they contained the needed information, did not have the form that was required for the project Taking as an example the memberships to clubs and professional associations it is demonstrated here how - with the help of content analysis - text data are converted into numeric data. For the analyzation of the text data the statistic program SPSS and the content analysis program INTEXT were used.

Dieser Beitrag soll zeigen, wie prozeßproduzierte Daten so umgewandelt werden, daß sie für eigene Analysezwecke brauchbar sind. Die Daten stammen aus dem Datensatz der Abgeordneten der Weimarer Republik (vgl. dazu Best 1991), dieser Datensatz liegt als SIR-Datenbank beim Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln vor und wird innerhalb eines Projektes über parlamentarische Eliten in Europa als Ausgangsdatensatz benutzt.

Für die Abgeordneten der deutschen Parlamente sollte ihre Zugehörigkeiten zu Vereinen oder Verbänden analysiert werden. Dazu werden die Vereine und Verbände in Gruppen (Kategorien) zusammengefaßt werdea Mit dem Verfahren der computerunterstützten Inhaltsanalyse sollte dann aus diesen Namen die

*** Address all communications to Harald Klein, Friedrich-Schiller-Universität, Institut für Soziologie, D-07740 Jena.**
Überarbeitete Fassung eines Vortrages, den der Verfasser anläßlich des ZHSF-Workshops: 20 Jahre QUANTUM, vom 4.-7. Oktober 1995 in Köln, im Rahmen der Sektion 'Computergestützte Inhaltsanalyse am Beispiel historischer Textkorpora', gehalten hat.

Kategorisierung in Gruppen durchgeführt werden. Jeder Name bekommt einen Code, der angibt, zu welcher Gruppe dieser Verein oder Verband gehört. Die Software für computerunterstützte Inhaltsanalyse - benutzt wurde **LNTEXT** - schreibt dann diesen Code, wenn sie den Namen gefunden hat, in die Ausgabedatei, die dann statistisch weiter verarbeitet werden kann.

In den bereits vorhandenen Datensätzen waren die Bezeichnungen dafür bereits erfaßt. Zuerst stellten sich technische Hindernisse in den Weg, die historische Ursachen haben. Die Bezeichnungen waren mit einer älteren Version des Statistikprogramms **SPSS** erzeugt worden, die nur eine maximale Länge von 4 Zeichen je alphanumerischer Variable und Großbuchstaben zuließ. Die Bezeichnungen waren deshalb nicht in einer Variablen, sondern in mehreren Variablen gespeichert Daher ergeben sich die folgenden Arbeitsschritte:

- Extrahieren der relevanten Variablen aus dem **SPSS**-Datensatz
- Zusammenfügen der Vereins- und Verbändenamen
- Entwickeln eines Kategoriensystems mit Suchbegriffen
- Codierung der Suchbegriffe
- Kontrolle des Codiervorgangs und Weiterentwicklung des Kategoriensystems
- Aggregierung der Daten zum Ändern der Fallbasis

Für das Extrahieren der Variablen wurde der **SPSS**-Befehl **LIST VARIABLES** benutzt, zwischen den Variablen stehen Leerzeichen, die aus der **SPSS**-Ausgabedatei entfernt werden müssen. Dies wurde mittels eines Makros von WordPerfect erledigt. Abbildung 1 zeigt die Daten, wie sie mit **LIST VARIABLES** erzeugt wurden. Vor den Bezeichnungen stehen noch drei numerische Variablen, die später für die statistische Weiterverarbeitung noch gebraucht werden.

Nach dem Entfernen der Leerzeichen sehen die Angaben der Vereine und Verbände dann wie in Abbildung 2 dargestellt aus.

Dieses Format der Daten kann **INTEXT** direkt verarbeiten. Die ersten 4 Zahlen werden als Identifikatoren benutzt, sie repräsentieren Variablen des Datensatzes: Identifkationsnummer des Abgeordneten und Angaben über Art und Dauer der Vereins- bzw. Verbandstätigkeit. Diese Variablen sind notwendig, weil nur die Tätigkeit vor Antritt eines Mandats in Führungspositionen gebraucht wird. **INTEXT** erwartet beim festen Format, daß die Identifikatoren und der Text immer auf den gleichen Spalten stehen.

Aus diesen Daten erzeugt **INTEXT** dann eine sogenannte Systemdatei, die Basis jeder weiteren Verarbeitung ist. Aus der Systemdatei entsteht dann eine Wörterliste, ein alphabetisch sortierte Verzeichnis aller im Text vorkommenden Zeichenketten (meistens Wörter, aber nicht immer). Die Wörterliste hat zwei Funktionen:

- Korrektur der Orthographie und Transkription
- Suchen von Zeichenketten, die als Suchbegriffe für die Kategorien geeignet sind, dazu werden Füllwörter (z.B. Artikel, Präpositionen usw.) aus der Wörterliste entfernt und die Wörterliste mehrspaltig ausgedruckt

Abbildung 1: Ausgabe der Variablen mit LIST VARIABLES

```
   6    5    7    3 GEWE RKSC HAFT  CHR ISTL ICH
   1    9    2    1 JUNG DT P RESS EDIE NST
   2    4    2    2 LAND GEME INDE TAG  EINZ ELST AAT
   3    1    6    2 ANWA LTSK AMME R
   4    3    7    2 DT A RBEI TERB UND
   5    9    6    3 SPD
1791    2    4    1 KPD  BEZI RK
   7    0    0    0
   8    9    8    1 GENO SSEN SCHA FT
   9    9    1    2 VB D ER H AUSA NGES TELL TEN
  10    5    2    2 DTVO ELKI SCHE  FRE IHEI TSPA RTEI
  11    0    0    0
  12    2    7    2 BAUE RNVE REIN  UNT ERFR ANKE N CH RS
  17    9    2    4 JUGE NDBE WEGU NG P ROLE TARI SCH
  14    5    2    4 REIC HSKA MMER BUND
  15    9    6    4 SCHN EIDE RVER BAND
  16    2    7    1 NSDA P KR EIS
  13    1    1    4 BURS CHEN SCHA FT
```

Abbildung 2: Antworten ohne überzählige Blanks

```
   6    5    7    3 GEWERKSCHAFT CHRISTLICH
   1    9    2    1 JUNGDT PRESSEDIENST
   2    4    2    2 LANDGEMEINDETAG EINZELSTAAT
   3    1    6    2 ANWALTSKAMMER
   4    3    7    2 DT ARBEITERBUND
   5    9    6    3 SPD
1791    2    4    1 KPD BEZIRK
   7    0    0    0
   8    9    8    1 GENOSSENSCHAFT
   9    9    1    2 VB DER HAUSANGESTELLTEN
  10    5    2    2 DTVOELKISCHE FREIHEITSPARTEI
  11    0    0    0
  12    2    7    2 BAUERNVEREIN UNTERFRANKEN CHRS
  17    9    2    4 JUGENDBEWEGUNG PROLETARISCH
  14    5    2    4 REICHSKAMMERBUND
  15    9    6    4 SCHNEIDERVERBAND
  16    2    7    1 NSDAP KREIS
  13    1    1    4 BURSCHENSCHAFT
  18    1    1    1 METALLARBEITERVERBAND CHRISTL
  19    9    2    4 SPARTAKUSBUND
  20    1    4    1 ZENTRUM
  21    1    6    1 GRDT FOEDERALIST HAMBG RECHTSP
  22    1    2    2 NSDAP
```

Dabei tauchen Transkriptionsprobleme auf. Wie schon aus Abbildung 2 erkennbar, wurden Abkürzungen verwendet. Das geschah aber nicht immer einheitlich, und so mußten die Texte erst einmal standardisiert werden. Die Standardisierung der Schreibweisen ist auch für die Formulierung von Suchbegriffen wichtig. Bei fehlender Standardisierung muß sonst jede Schreibweise eines Vereines oder Verbandes einzeln aufgeführt werden, und das ist nicht nur zeitlich aufwendig, sondern auch fehleranfällig. Nachdem die Vereins- und Verbandsnamen standardisiert worden waren, wurden sie in Kategorien zusammengefaßt (Abbildung 3).

Abbildung 3: Kategorien der Vereine und Verbände

Code	Bedeutung
1	Gewerkschaften
2	Landwirtschaft
3	Industrie
4	Mittelstand
5	Freiberufler
6	Kultur
7	Kirchen
8	völkisch
9	Jugend
10	Sport
11	Geselligkeit
12	Partei
13	soziale Sicherung
14	Beamtenorganisationen
15	Karitas
16	Abstinenzler
17	Studenten
18	Akademien
19	Aufsiehtsräte
20	Genossenschaft
21	Bildung
22	Krieger
23	Reform
24	Frauen
25	Frieden
26	Akademiker
27	Betriebsräte
28	Feuerwehr
29	Heimat
30	Friedensgesellschaften
31	Jüdische
32	Kommunal
33	Stände
34	politisch
36	Mitteleuropa
37	Presse
38	Reichsbanner
39	Mieter
40	Technik
41	linksrevolutionär
42	unspezifisch

Für jede Kategorie werden anhand der Wörterliste Suchbegriffe definiert. Wird ein Suchbegriff gefunden, dann wird dessen Code in die Ausgabedatei geschrieben und kann statistisch weiter verarbeitet werden, INTEXT schreibt diese Codes als **ASCII**-Datei in Form einer Rechtecksmatrix (vgl. Abbildung 5) heraus und generiert auch die passenden **sPSS**-Befehle (DATA LIST und VARIABLE LABELS). Jede Antwort gilt als ein Fall, und da maximal 7 Vereins- und Verbandszugehörigkeiten vorgesehen sind, gibt es für jeden Abgeordneten 7 Fälle, die allerdings nicht immer besetzt sind. Bei der Definition der Suchbegriffe gibt es mehrere Möglichkeiten;

- Wortteile: Wortanfänge, Wortendungen, Zeichenketten (Strings)
- ganzes Wort
- mehrere Wörter, z.B. Namen (Hans-Jochen Vogel, Bernhard Vogel)
- Wortstamm folgen, z.B. gut) (politik)

Abbildung 4: Beispiele für die Suchbegriffe

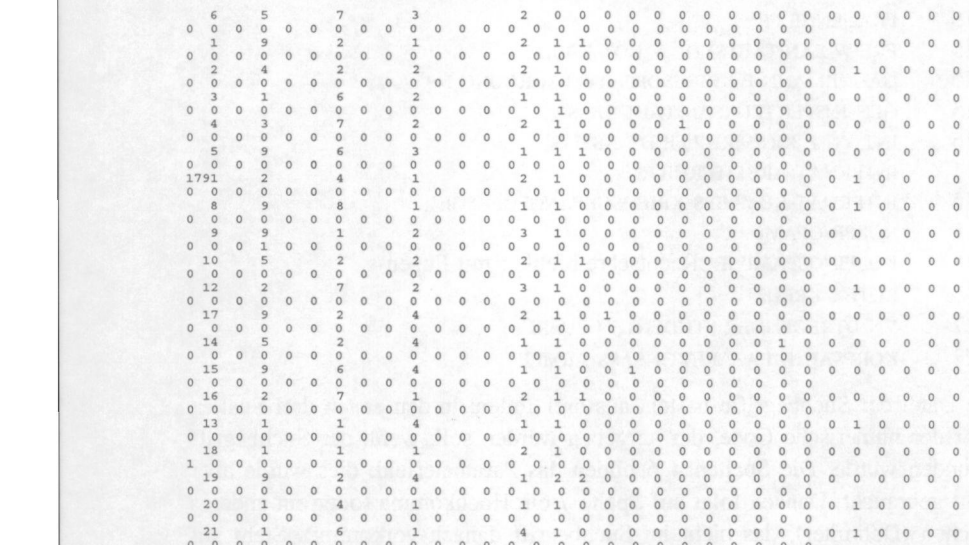
```
1   ' ANGESTELLTER DT TAPEZIERERVERB'
1   ' ARBEITERVERTRETERVER DRESDEN'
1   ' ARBSPORTBEWEGUNG'
1   ' ARBVEREINSBEWEGUNG'
1   ' KATH ARBINTERNATIONALE'
1   ' KATH GESELLENVER'
1   ' KATH GESELLVNBEREIN'
1   ' KATH LEHRER' INNENVER KOELN
1   ' LEHRERRAT'
1   ' LEHRERVER'
1   ' R BD DT EISENB VORSTEHER U SEKRETAER'
1   ' R BD VATERLD ARB U WERKVEREINE VORS'
2   ' PROV VERBD RASSEZIEGENZUECHTER'
2   ' REICHSARBGM LW FORSTARBGEBERVB'
2   ' REICHSBD AKADEMISCH GEBILDETER LWIRT'
2   ' REICHSVB DT LW G'
2   ' REICHSVB LW KLEINBETRIEBE'
3   ' DT WIRTSCHAFTSVB F SUED MITTELAMRKA'
3   ' EISEN U STAHLWAREN INDUSTRIEBD'
3   ' HANSA BUND'
3   ' HANSABUND'
4   ' DT TEXTILDETAILLISTENVB'
4   ' DT TEXTIL DETAILLISTEN VB'
6   ' ARCHAEOLOGISCHE GESELLSCHAFT BRUESSL'
6   ' BD D WERKLEUTE BD F SCHOEPF ARBEIT'
6   ' CARNEGIE STIFTUNG EUROP BUEREAU'
```

```
 7   ' EV> BUND'
 8   ' REICHSVERBAND Z BEKAEMPFUNG D SPD'
 9   ' HJ'
12   ' CHRIST SOZ> VOLKSDIENST'
15   ' BETREUNG KATH ITAL ARBEITER HANNOVER'
15   ' FRAUENHILFE'
15   ' FREIW SANITAETSKOL V ROT KREUZ'
15   ' GAUOBMANN F KRIEGSOPFERVERSORGUNG NS'
15   ' GEN INSPEKTEURSANITAET SASS'
15   ' INT VG F KINDERHILFE DT ABTLG'
 1   ' INTERNAT ARBEITERHILFE'
15   ' INTERNAT KOMMISS KRIEGSGEFANGFUERSGE'
15   ' KOLPINGFAMILIE'
15   ' KOLPINGSFAMILIE' Rechtschreibfehler: mit Fugen-s
15   ' ROTE> KREUZ '
17   ' VN DTREDENDER STUDENTEN ZUERI'
22   ' KORPSAU D LAZARETTINSASS ARMEE'
```

Die Datei der Suchbegriffe besteht aus drei Teilen: in den ersten drei Spalten steht der numerische Code, der vergeben werden soll, wenn der Suchbegriff gefunden wurde. Die Spalten 4-6 bilden das Parameterfeld, dies wurde aber nicht gebraucht. Danach folgt auf Spalte 7 ein Hochkomma (oder ein anderes Zeichen (Delimiter), das nicht im Suchbegriff danach vorkommt). Steht auf Spalte 8 ein Leerzeichen, so handelt es sich beim Suchbegriff um ein ganzes Wort oder einen Wortanfang, steht dort kein Leerzeichen, kann es sich um eine Zeichenkette handeln, die an beliebiger Stelle innerhalb eines Wort (Anfang, Mitte oder Ende) vorkommen darf. Der Suchbegriff ' LEHRERRAT' findet einen Vereins- oder Verbandsnamen LEHRERRAT, aber nicht GYMNASIALLEHRERRAT, wohl aber LEHRERRATSMITGLIED. Beim Suchbegriff ' ROTE> KREUZ ' handelt es sich um eine Wortstammfolge, mit der Textstücke wie ROTES KREUZ oder ROTEN KREUZ gefunden werden (vgl. dazu Klein 1995, S. 96). Nach dem zweiten Delimiter kann auch ein Kommentar folgen, der nicht mit verarbeitet wird.

Bei der Durchsicht der Wörterliste ist es eine effiziente Arbeitsweise, die Suchbegriffe möglichst so zu definieren, daß möglichst wenige gebraucht werden. Dies dient zum einen der Übersichtlichkeit des Kategoriensystems und zum anderen der Geschwindigkeit der Codierung. Vor der Codierung müssen nur noch die Kategorienetiketten definiert werden. Diese Datei sieht etwa so aus wie Abbildung 2; in jeder Zeile stehen Code und dessen inhaltliche Bedeutung. Die Kategorienetiketten zwingen nicht nur zur Dokumentation des Kategoriensystems, sondern werden bei der interaktiven Codierung und bei der Erzeugung der **SPSS**-Befehle benutzt

Abbildung 5: Ergebnisse der Codierung



Je kürzer die Suchbegriffe werden, desto eher können sie mehrdeutig werden. BAUERN ist zwar in den meisten Fällen der Landwirtschaft eindeutig zugeordnet, doch bei der BAUERNBANK erfolgt eine Fehlcodierung. Um dies unter Kontrolle zu bringen, mußten die kurzen Suchbegriffe mit KWICS (Key-Word-In-Context) Listen auf Eindeutigkeit kontrolliert werden. Dabei stellte sich heraus, daß eine erhebliche Anzahl von Suchbegriffen zu Fehl- und Mehrfachcodierungen geführt hätte. Das ursprüngliche Kategoriensystem mußte erweitert werden. Nach der erfolgten Codierung mußte festgestellt werden, ob auch alle Antworten zu einer Codierung führten. Das war zuerst nicht der Fall, deshalb wurden anhand der nichtcodierten Texteinheiten die Antworten herausgesucht, die uncodiert geblieben waren. Oft waren es Rechtschreibfehler oder ungewöhnliche Abkürzungen. Durch Änderungen der Antworten und der Suchbegriffe wurden dann alle Antworten codiert, und somit standen die Daten für die statistische Analyse zur Verfügung.

Ein weiteres Problem war die Mehrfachcodierung von Antworten. Ein gutes Kategoriensystem sollte keine mehrfachen Codierungen zulassen, INTEXT bietet mit der Dublettenkontrolle eine Vermeidungsstrategie an, die darauf beruht, daß ein Test gemacht wird, ob es Suchbegriffe gibt, die in anderen enthalten sind. Dabei wird auch getestet, ob in Teil einer Wortstammfolge in einem anderen Suchbegriff vorkommt. Das Ergebnis wird in eine Datei geschrieben und am Bildschirm angezeigt, das Kategoriensystem wird dabei aber nicht geändert; Änderungen muß man selbst vornehmen, indem man Suchbegriffe löscht, ändert oder hinzufügt. Bei der Codierung selbst werden Mehrfachcodierungen zur Zeit noch durchgeführt, ein Eingreifen oder Steuern ist aber mit der interaktiven Codierung möglich, wenn auch aufwendig. Als Weiterentwicklung ist denkbar, daß bereits codierte Textstellen gesperrt werden, daß der jeweils letzte Suchbegriff gilt oder daß interaktiv gearbeitet wird.

Die numerischen Ergebnisse zeigt Abbildung 5. Die ersten vier Variablen sind die drei Identifikatoren, es folgen die Anzahl der Wörter und die Anzahl der Codes, danach die Zähler für die insgesamt 42 Kategorien. Im zweiten Fall wurde Code 1 einmal vergeben.

Bibliographie:

Best, Heinrich; Ralph Ponemereo (1991): The German Parliamentarian Data Base. Catching the Complexities of Political Life-History. In: Best, Heinrich u.a. (Hrsg.): Computers in the Humanities and the Social Sciences, München.

Klein, Harald (1995): INTEXT 3.0 Handbuch. Jena.